Authors' Response to Comments of Reviewer #2

Referee comments are in black. Author responses are in blue and revised manuscript in blue marked by underline.

This study integrates ground observations, satellite remote sensing, reanalysis data, and machine learning techniques to systematically analyze the radiative forcing (RF) characteristics, influencing factors, and their relative importance to precipitation during severe autumn and winter PM2.5 heavy pollution events in the Bohai Sea region (2014–2023). While the work is meaningful, the following issues need to be addressed:

Response: We sincerely appreciate your valuable comments and thoughtful suggestions for providing us with important guidance for improving the manuscript. We have carefully addressed each of the issues you raised below with the following amendments.

1. Page 6, Line 135: The term "clean days" is used without a clear definition. Please specify the exact PM2.5 concentration threshold or criteria used to classify a day as "clean".

Response: Thank you for your suggestion. We have clarified the regional clean day in the section of data and method in the modified version: "The regional clean day is defined as the day with all stations $PM_{2.5} < 75\ \mu gm^{-3}$ within the study area".

2. Although machine learning algorithms are central to this study, the manuscript lacks quantitative performance evaluation. Please provide essential statistical metrics, such as the coefficient of determination (R^2), RMSE, and MAE, for both the radiative forcing and precipitation prediction models to validate their accuracy.

Response: We have incorporated your recommendation by adding an explanation of the statistical metrics for the machine learning models concerning radiative forcing and total daily precipitation in the section "2.4.2 Importance estimation based on Random Forest algorithms" in the revised manuscript as "The coefficient of determination (R^2),

root mean square error (RMSE) and mean absolute error (MAE) were used to evaluate the model performance, which was shown in Table S2.". Additionally, we have included the following Table 1 (Table S2 in the new supplementary file) in the supplementary materials:

Table 1. The essential statistical metrics of the evaluation of Random Forest training models.

Statistical metrics	RF surface clear	RF surface all	RF TOA clear	RF TOA all	RF Atmos clear	RF Atmos all	Total pre
R^2	0.6721	0.6722	0.6852	0.6512	0.6557	0.6742	0.5728
RMSE	0.5697	0.5697	0.5562	0.5881	0.5837	0.5683	0.6454
MAE	0.4277	0.4291	0.3921	0.4381	0.4442	0.4416	0.2410

It should be emphasized that this paper focuses solely on $PM_{2.5}$, temperature, and winds as factors. Consequently, the model's key statistical metrics are not particularly high. However, our objective is to compare the relative importance of $PM_{2.5}$, temperature, and winds, rather than the accuracy of predictive outcomes. Therefore, these model evaluation metrics have been placed in the supplementary file. Nevertheless, we recognize that the importance factors for radiation and precipitation derived through machine learning methods may be influenced by other unaccounted variables, such as terrain elevation and land surface conditions. Consequently, we added the machine learning model accuracy limitations and the shortcomings of excluding factors in the discussion, as "TAP $PM_{2.5}$ data contains the meteorological information and some other factors have not been taken into account in the machine learning such as terrain elevation. The mechanism and degree of the impact of pollution RF on precipitation are not unequivocal.".

3. Page 20, Line 455: The discussion regarding regional heterogeneity is currently superficial. The authors should elaborate on the underlying physical mechanisms driving these differences. Specifically, the analysis should link the results to sub-regional variations in surface types, emission intensities, and topographical features.
Response: According to your suggestion, we have added the discussion regarding regional heterogeneity in section 3.3.1 and section 4. In section 3.3.1, we discussed the reason for regional heterogeneity in depth. And in section 4, we also briefly explain the

regional heterogeneity. The adding content in Section 3.3.1 is as followed: "The importance of influencing factors to pollution RF showed heterogeneity across different regions. This may be related to variations in aerosol concentrations and pollution RF. In clear sky, the BLT region featured high $PM_{2.5}$ concentrations (Figure 3) coupled with high RF values (Figure 4a), whereas the SB region exhibited low $PM_{2.5}$ concentrations alongside low RF, illustrating the higher importance of $PM_{2.5}$ impacts on RF at the surface in these two regions (Figure 8a). Conversely, the NB region, with low $PM_{2.5}$ concentrations but high RF, showed less importance of $PM_{2.5}$ than in the BLT and SB. Differences in the importance of meteorological parameters across regions may be related to local topography and geographical location. Figure 1 indicates that the northwestern parts of the BLT and NB regions border mountainous terrain, where wind direction and speed significantly influence pollution dispersion and transport. Consequently, wind (V-wind in the BLT region, U- and V-winds in the NB region) exhibited greater importance for RF in these regions compared to the SB region. Conversely, the SB region's lower latitude and absence of nearby high mountain ranges resulted in temperature factors being more important than in the BLT and NB regions." And in the Section 4 with "The regional heterogeneity for the important factors in the Bohai Rim regions may relate to aerosol concentrations, RF values, local topography, and geographical location."

4. Please clarify whether data standardization was performed prior to K-means clustering. Given the disparate units and scales of the input variables, omitting standardization could significantly bias the clustering results towards variables with larger magnitudes. If standardization was not applied, a rigorous justification is required.

Response: For k-means clustering, we did not perform data standardization. This is because our clustering is conducted separately for meteorological parameters, where there is no issue of inconsistent units. Furthermore, we selected parameters below 850 hPa for clustering, meaning the data used for clustering exhibited minimal variation in magnitude. Concurrently, in response to your comments, we performed standardization prior to clustering. The results were largely consistent with those obtained without standardization, as illustrated in Figure 1 below. Regarding this standardization issue, we have included the figure in the supplementary file and provided an explanation in

the section "2.4.3 The data usage workflow" as "Since we cluster based on separate parameters in the boundary, the standardization before clustering becomes unnecessary, as standardized and unstandardized results are essentially consistent (see Figure S3)".

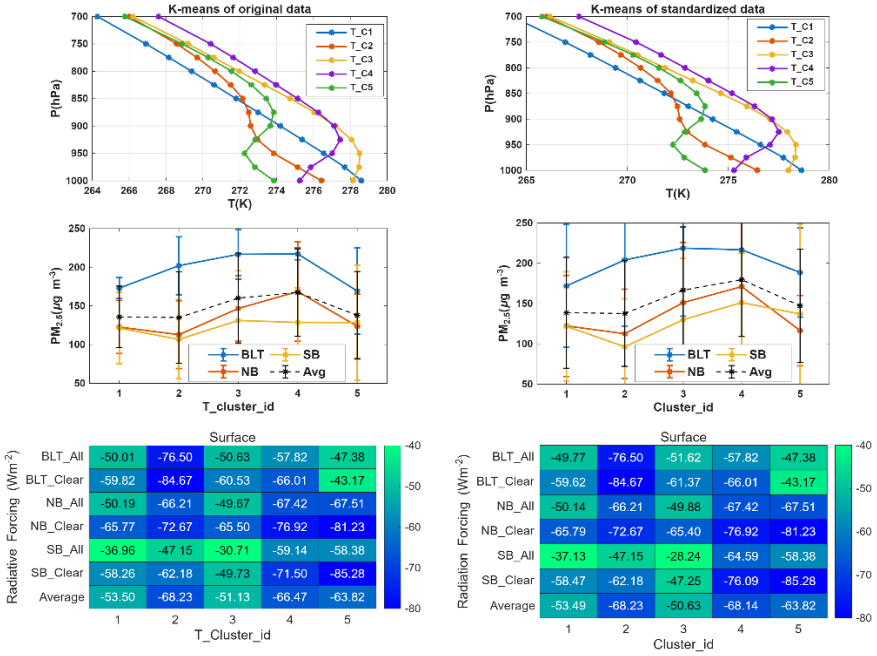

Figure 1. The comparison of k-means clustering by using original data and standardized data (Z-score of T). (Figure S3 in new Supplementary file).

5. Data Quality and Pre-processing: The validity of applying machine learning to linearly interpolated CERES radiation data requires justification. Linear interpolation may introduce artifacts or smooth out extreme values, potentially biasing the training process of the machine learning models. The authors should discuss: The extent of missing data prior to interpolation. Whether the spatial resolution of CERES data is sufficient to capture the local variability of pollution events in the study area. How these uncertainties affect the generalizability of the conclusions.

Response: We agree with your viewpoint.

①Regarding the extent of missing data prior to interpolation, we used the CERES-SYN1deg product; "CERES-SYN1deg provides hourly gridded observed top of atmosphere (TOA) fluxes and computed surface fluxes from the Fu-Liou radiative transfer model, which is suitable for regional diurnal and process studies. The CERES-SYN1deg products have been validated by other measurements (Doelling et al., 2016;

Fillmore et al., 2022; Rutan et al., 2015)". Our work utilizes the daily numerical data, thus eliminating the issue of missing data prior to interpolation.

②Regarding whether the spatial resolution of the CERES data can capture local variability in pollution events within the study area, it can be assessed from Figure 2 (Figure S4 in the new supplemental file). The figure demonstrates that the raw 1x1° data exhibits differences between the three study regions (BLT, SB, and NB), with some local variability also present within each region. Naturally, the CERES SYNIdeg 1°× 1° spatial resolution data cannot resolve minute gradients within urban areas. However, this study focuses on regional radiative forcing and its influencing factors, with the temporal focus being regional heavy pollution days (regional pollution events). Regional heavy pollution days typically exhibit strong spatial consistency across larger scales. Therefore, this dataset can still capture regional variations in pollution events to a certain extent in study regions on the regional heavy pollution days.

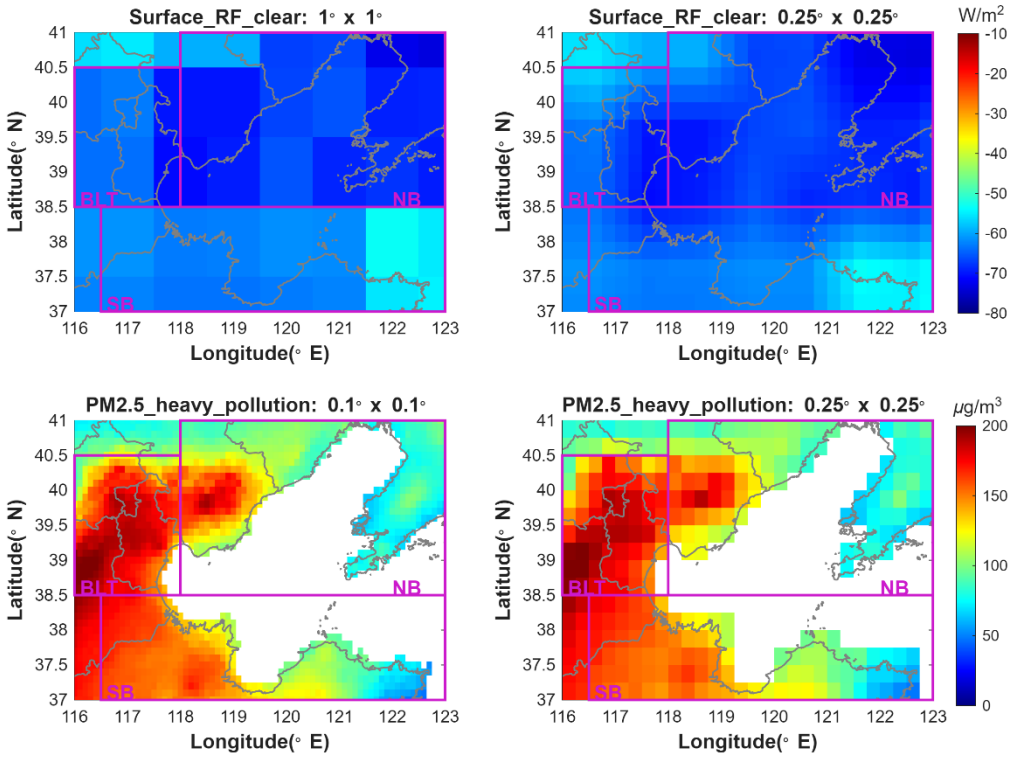

Figure 2. The comparison between the original data (first column) and the interpolated data (second column) in the study area. The first row is the radiative forcing (RF) from CERES at the surface in clear sky, and the second row is the interpolation for the $PM_{2.5}$ from TAP. The purple rectangle indicates the three regions.

③ Concerning the bias issues arising from machine learning following interpolation, we can find in Figure 2 that "The spatial distributions of the interpolated and original datasets are generally consistent, with minor discrepancies observed only at a few grid points exhibiting abrupt value changes. Since the results of this study primarily focus on regional averages, the errors introduced at a limited number of grid points have less impact on the regional mean outcomes." Besides, we have provided an explanation in the discussion section of the paper. Due to issues with data resolution, higher-precision observational data will be required in the future, as "The interpolation of datasets with different spatial resolutions used for training of the machine learning algorithm may cause some uncertainty. Further studies are required to quantify the impacts of these factors on RF and precipitation and to explore the underlying physical mechanisms or connections through additional observations in diverse regions and in higher spatial resolution, and numerical simulations."

6. In Section 2.4 ： The current description of K-means and Random Forest is too generic. This section should be condensed to focus on the specific implementation details specific to this study, such as hyperparameter settings, input feature selection, and the cross-validation strategy employed.

Response: Thanks for this suggestion. We have condensed the description of the section about K-means and Random Forest and added some specific implementation details to this study including hyperparameter settings, input feature selection, and the cross-validation strategy employed. Besides, we added a section of "2.4.3 The data usage workflow" and Figure 3 to show more implementation details.

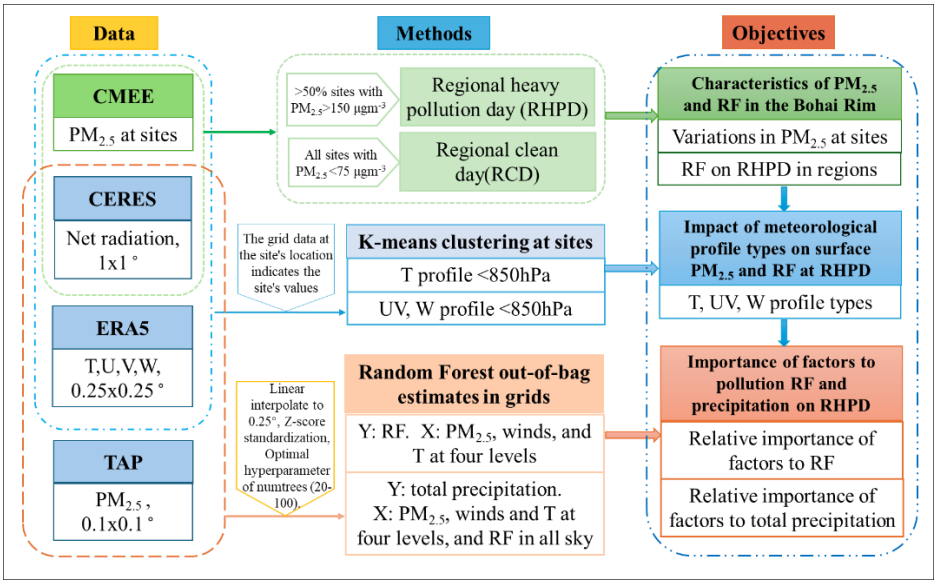

Figure 3. The data usage workflow and framework in this work. (Figure 2 in the revised manuscript).

For k-means, the description is as "We classified the T and wind components profiles in the boundary layer (below 850 hPa) adopting the k-means clustering method (Lloyd, 1982). K-means clustering is an unsupervised machine learning algorithm used to partition a dataset into distinct groups or clusters, and popular in a wide variety of applications due to its simplicity, efficiency and effectiveness. Through calculation, there were 161 days of regional $PM_{2.5}$ heavy pollution days in the study regions, which is shown in the next section. Then, the profiles at the 11 stations in Bohai Rim on the 161 regional heavy polluted days are used to cluster (the number of samples is 11*161). We use the k-means ++ algorithm for cluster center initialization (Arthur and Vassilvitskii 2006) and the squared Euclidean distance to measure the similarity to the centroid. We selected the numbers of clusters (2–8) for classification of T, and then combined the elbow method (the corner of the Sum Square Error) and the representativity of T profiles to determine the last number of clusters (=5 in this study). The numbers of horizontal and vertical wind component clusters (also 5 clusters) were selected along the T clustering."

For Random Forest, the description is "We used the Random Forest algorithm to compare and rank the importance of various factors to pollution RF and daily total precipitation. The variable factors concerned in this study were $PM_{2.5}$, T, and 3 wind components at four levels (500, 700, 850, and 1000 hPa). The Random Forest method is a popular ensemble learning technique that combines multiple decision trees to

improve prediction accuracy and reduce overfitting (Breiman, 2001). In addition, it performs excellently for evaluating the independent variables' importance (Cutler et al., 2007). This study mainly used the "out-of-bag" observations method (Archer and Kimes, 2008) in the Random Forest regression model to calculate the importances of variables. Out-of-bag predictor importance estimates by permutation measure how influential the model's predictor variables are at predicting the response. Thus, the larger the calculated value, the greater its importance. For the random forest model training, this study employed a widely used 10-fold cross-validation (CV) method. Through repeated tests, we obtained the optimal hyperparameter of the number of trees from 20 to 100 and the number of leaf used the default of 5. The coefficient of determination (R^2), root mean square error (RMSE) and mean absolute error (MAE) were used to evaluate the model performance, which was shown in Table S2."